

Fast Product-Matrix Regenerating Codes

Nicolas Le Scouarnec
Technicolor, France

Abstract

Distributed storage systems support failures of individual devices by the use of replication or erasure correcting codes. While erasure correcting codes offer a better storage efficiency than replication for similar fault tolerance, they incur higher CPU consumption, higher network consumption and higher disk I/Os. To address these issues, codes specific to storage systems have been designed. Their main feature is the ability to repair a single lost disk efficiently. In this paper, we focus on one such class of codes that minimize network consumption during repair, namely regenerating codes. We implement the original Product-Matrix Regenerating codes as well as a new optimization we propose and show that the resulting optimized codes allow achieving 790 MB/s for encoding in typical settings. Reported speeds are significantly higher than previous studies, highlighting that regenerating codes can be used with little CPU penalty.

1 Introduction

The increasing needs for storage in datacenter pushed by numerous cloud-based storage services has lead to the development of storage systems with a better tradeoff between reliability and storage. The usual solution is to rely on erasure correcting codes [3, 6, 8, 9, 22]. While they allow decreasing storage costs, they generally come with higher costs in term of network, I/O or CPU and thus hardware cost and power consumption. To alleviate these drawbacks, optimized codes have been designed: (i) regenerating codes minimize network-related costs [5, 6, 23], (ii) locally repairable codes minimize I/O related costs [10, 12, 14, 17], (iii) other codes minimize CPU related costs [16, 18, 21]. In the rest of the paper, we will focus on regenerating codes as they offer the best tradeoff between network and storage.

Regenerating codes have been mainly studied with respect to either theoretical aspects (*i.e.*, existence) [2,

5, 6, 23–29]; or cost in bandwidth [4, 11, 15]. Beside these studies, few have looked at system aspects, including encoding/decoding throughput (*i.e.*, CPU cost). The throughputs reported are 50 KB/s ($k=32$, $d=63$) in [7], 0.6 MB/s ($k=16$, $d=30$) in [13], 100MB/s ($k=2$, $d=3$) [4]. So far, reported speeds, except for very low values of k , are incompatible with practical deployments.

In this paper, we go beyond these, and provide insights useful to practical deployment of regenerating codes.

- We describe an optimization that almost quadruple the performance of product-matrix codes [23] (*e.g.*, from 210 to 790 MB/s for $k = 8$ systematic codes)
- We report throughputs for product-matrix regenerating codes and compare them to Reed-Solomon codes. For systematic codes with $k = 8$, our optimized product-matrix codes encode at 790 MB/s when Reed-Solomon codes encode at 1640 MB/s.

Section 2 describes some background on product-matrix regenerating codes as well as libraries we rely on for our implementation (*i.e.*, Jerasure [20] and GF-Complete [19]). Section 3 describes the transformations we apply to product-matrix codes to turn them into linear codes, or to enhance the performance of the systematic form by sparsifying the encoding matrix. Section 4 shows the performance achieved and studies the impact of the various parameters.

2 Background

Fault tolerance mechanisms such as replication or erasure correcting codes are used to limit the impact of failures in distributed storage systems. An erasure correcting code encodes k blocks of original data (column vector X) to n blocks of encoded data (column vector Y), so that the k original blocks can be recovered from any k encoded blocks. The code can be defined by its generator matrix G of dimension $n \times k$. The encoding operation is

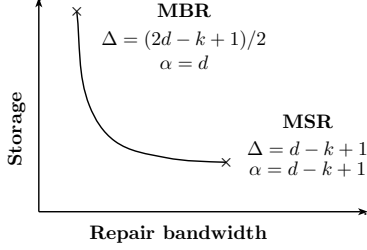


Figure 1: The tradeoff curve of regenerating codes with Minimum Storage Regenerating codes (MSR) and Minimum Bandwidth Regenerating codes (MBR)

then expressed as $Y = GX$; and the decoding operation as $X = \tilde{G}^{-1}\tilde{Y}$ where \tilde{G} (resp. \tilde{Y}) are the rows of G (resp. Y) that correspond to the remaining encoded blocks. Each device stores one block of data.

A key feature of distributed storage systems is their ability to self-repair (*i.e.*, to recreate lost encoded blocks whenever some device fails). Erasure correcting codes support such repair by decoding k blocks and encoding again. This implies the transfer of k blocks to recreate a single lost block, leading to high network-related repair cost. To solve this issue, regenerating codes [5] have been designed as randomized codes offering an optimal tradeoff between storage and network bandwidth (*i.e.*, network-related repair cost). Then, numerous studies [2, 6, 24–29] have focused on deterministic regenerating codes including product-matrix codes which are applicable to a wide set of parameters [23].

Regenerating codes achieve the optimal tradeoff between storage and bandwidth with two main variants: (i) MSR codes that favor storage, and (ii) MBR codes that favor bandwidth, as depicted on Figure 1. MSR codes encode $k\Delta$ blocks to $n\alpha$ blocks with $\alpha = \Delta = d - k + 1$. Each device stores α blocks. During repair, d blocks are transferred over the network. MBR are similar but have $\alpha = d$ and $\Delta = (2d - k + 1)/2$ thus resulting in higher storage costs but lower network bandwidth usage.

Regenerating codes can be implemented using linear codes. Similarly to regular erasure correcting codes, the code can be defined by its generator matrix G of dimension $n\alpha \times k\Delta$. The encoding operation is then expressed as $Y = GX$ (where X is a column vector of $k\Delta$ blocks and Y is a column vector of $n\alpha$ blocks); and the decoding operation as $X = \tilde{G}^{-1}\tilde{Y}$ where \tilde{G} (resp. \tilde{Y}) are the rows of G (resp. Y) that correspond to the remaining encoded blocks. The factor α in dimensions comes from the fact that each device stores α instead of 1 block for erasure correcting codes¹. As a consequence, the encod-

¹This allows devices using regenerating codes to repair α blocks by downloading 1 block from d devices thus saving bandwidth when compared to erasure correcting codes that repair 1 block by downloading 1 block from k devices.

ing and decoding complexities come with an additional factor $\alpha\Delta$. Given that in a typical setting $\alpha \propto k$ and $\Delta \propto k$, the encoding and decoding complexities, which are $\Omega(k^2)$ for erasure correcting codes, become $\Omega(k^4)$ for regenerating codes.

Product-matrix regenerating codes [23] rely on specific encoding and decoding algorithms and are not defined using a generator matrix as usual linear codes. The $k\Delta$ original blocks of X are arranged (with repetitions and zeros) into a message matrix M of dimension $d \times \alpha$ in a way specific to MSR or MBR. The code is defined by its encoding matrix Ψ of dimension $n \times d$. The encoding operation is thus defined as $C = \Psi M$ such that C is a $n \times \alpha$ matrix containing the $n\alpha$ encoded blocks. The decoding algorithms are specific to the type of regenerating code used (MSR or MBR) and are described in [23].

Regenerating codes have been mainly studied for their saving in bandwidth, leaving aside computational performance. A few papers have implemented and measured the performance of various regenerating codes such as randomized codes in [4, 7, 13]. They achieve around 50 KB/s using a pure Java-based implementation ($k=32$, $d=63$) in [7, 13], 100MB/s for very small codes ($k=2$, $d=3$) in [4]. Implementation of deterministic codes achieves relatively higher speed such as 0.6 MB/s ($k=16$, $d=30$) in pure Java [13]. Encoding throughput as reported in these publications is a factor that is likely to limit the deployment of regenerating codes, since saving network bandwidth increase a lot processing costs.

Regarding the implementation of erasure correcting codes, highly optimized libraries have been recently released and provide fast arithmetic on Galois Field, namely GF-Complete [19] or linear codes, namely Jerasure 2.0 [20]. We will use these two libraries in our implementation of product-matrix regenerating codes.

3 Fast PM Regenerating codes

We now describe the key aspects of our implementation of product matrix regenerating codes. Basically, the best performance is achieved by turning the PM codes into systematic linear codes, and using the sparse code we define in this paper instead of the *vanilla* code from the seminal paper [23]. The performance impacts of choosing each optimization are detailed in Section 4.

3.1 Linearization of Product-Matrix codes

When implementing product matrix regenerating codes, two alternatives are possible, namely (i) applying the algorithms described in the seminal paper [23] or (ii) transforming them to linear codes and using generic algorithms for linear codes such as the ones implemented in

Jerasure [20]. To support our evaluation of both strategies in Section 4, we explain how product matrix codes can be transformed into equivalent linear codes.

To obtain the linearized version of the product matrix code, we need to create a $n\alpha \times k\Delta$ generator matrix G for a given code. First, we construct an index matrix L accordingly to the definition of the message matrix M of the seminal paper, so that $M_{i,j} = X_{L_{i,j}}$. For example, for an MSR code with $k = 3, \alpha = \Delta = 2$, we would have the index matrix L of dimension $2\alpha \times \alpha$ such that

$$L = \begin{bmatrix} 1 & 2 & 4 & 5 \\ 2 & 3 & 5 & 6 \end{bmatrix}^t$$

Encoding by the product-matrix regenerating code is defined as $C = \Psi M$. The element $C_{i,j}$ and the corresponding element $Y_{\alpha i+j}$ for linear regenerating code defined as $Y = GX$ are computed using

$$C_{i,j} = \sum_{l=1}^n \Psi_{i,l} X_{L_{l,j}} \quad Y_{\alpha i+j} = \sum_{l'=1}^{k\delta} G_{\alpha i+j,l'} X_{l'}$$

For all i, j , we have $C_{i,j} = Y_{\alpha i+j}$. Applying a change of variable $l' = L_{l,j}$, and noticing that (i) by construction (see [23]) of M and L , no row nor column of L has duplicates values, (ii) the equality can hold only if $G_{\alpha i+j,l'} = 0$ for any $X_{l'}$ not present on the left-hand side, we obtain

$$\sum_{l=1}^n \Psi_{i,l} X_{L_{l,j}} = \sum_{l'=1}^{k\delta} G_{\alpha i+j,L_{l,j}} X_{L_{l,j}}$$

This implies that the generator matrix G such that

$$G_{\alpha i+j,l'} = \begin{cases} \Psi_{i,l} & \text{if there exist } l \text{ such that } l' = L_{l,j} \\ 0 & \text{otherwise} \end{cases}$$

is equivalent to a product matrix code (MSR or MBR) defined by Ψ and L .

We explore the benefits of linearizing product matrix codes in Section 4, and see that it leads to significant improvement when using systematic codes. Also, as side advantages, linearization allows (i) re-using libraries designed for regular erasure codes thus significantly, (ii) simpler *zero-copy/lazy* (i.e., decoding only lost blocks) implementation as it is a single step transformation.

3.2 Systematic codes

A key feature of codes for storage, being erasure correcting codes or regenerating codes, is their ability to be transformed into systematic codes. A systematic erasure correcting code (resp. systematic regenerating code) is a code such that the k (resp. $k\Delta$) first blocks of Y are equal to the original data X . Systematic codes have two main advantages: (i) accessing data does not require decoding provided that none of the k first devices has failed

thus enhancing the performance of the common case², (ii) encoding tends to be less costly as we only need to compute the $n - k$ (resp. $n\alpha - k\Delta$) last rows of Y , the k (resp. $k\Delta$) first being equal to X by construction.

The construction of systematic codes is based on the fact that encoding and decoding operations are more or less commutative. Hence, one can apply the decoding to the original data X to obtain precoded data Z and then encoding Z to obtain Y such that the $Y_{[1 \dots k\alpha]} = X$. For linear codes, the encoding is thus defined as $Y = G\tilde{G}^{-1}X$ with \tilde{G} being the $k\Delta$ first rows of G . The resulting systematic code has a generator matrix $G' = G\tilde{G}^{-1}$, such that $G' = [I \quad G''']^t$. By construction, G' inherits appropriate properties (e.g., all $k \times k$ (resp. $k\alpha \times k\alpha$) sub-matrices are full-rank) from G ensuring that it is an appropriate generator matrix. An alternative could be to choose G'' such that $G' = [I \quad G''']^t$ has the needed properties.

While building G' directly is possible for some codes, this is not easy for MSR product-matrix regenerating codes since matrix Ψ (from which G can be derived as explained in Section 3.1) must satisfy several constraints to allow efficient repair. Hence, the seminal paper builds systematic MSR codes by successively decoding and encoding. For MBR product-matrix codes, since the matrix Ψ is less constrained, the seminal paper [23] directly gives a systematic encoding matrix $\Psi = [I \quad \Psi''']^t$ thus leading to simpler efficient implementation.

However, as we explore in Sub-section 3.3 and Section 4, such indirect construction of Product-Matrix MSR codes has a significant impact on the computing cost thus requiring optimization beyond the construction from paper [23] as presented in the next sub-section.

3.3 Sparse MSR PM codes

The paper [23] specifies the following constraints on the encoding matrix Ψ for MSR product-matrix codes.

- $\Psi = [\Phi \quad \Lambda\Phi]$ where Φ is a $n \times \alpha$ matrix and Λ is a $n \times n$ diagonal matrix.
- Any d rows of Ψ are linearly independent.
- Any α rows of Φ are linearly independent.
- All values of Λ are distinct.

The construction suggested in the paper is to take Ψ as a Vandermonde matrix (i.e., $\Phi_{i,j} = g^{(i-1)(j-1)}$) that satisfies all needed properties. For example, if $n = 5, k = 3, d = 4, \alpha = 2$ and the finite field used has a generator element g , this gives the following matrices:

$$\Phi = \begin{pmatrix} 1 & 1 \\ 1 & g^1 \\ 1 & g^2 \\ 1 & g^3 \\ 1 & g^4 \end{pmatrix} \quad \Lambda = \begin{pmatrix} 1 & & & & \\ & g^2 & & & \\ & & g^4 & & \\ & & & g^6 & \\ & & & & g^8 \end{pmatrix} \quad \Psi = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & g^1 & g^2 & g^3 \\ 1 & g^2 & g^4 & g^6 \\ 1 & g^3 & g^6 & g^9 \\ 1 & g^4 & g^8 & g^{12} \end{pmatrix}$$

²Even if the system is designed to tolerate failures, most of the time, the device storing the data accessed is likely to be available.

Table 1: Sparsity (*i.e.*, percentage of 0) of the generator (G for non-systematic and G'' for systematic).

Code (n=2k-1, d=2k-2)	k=4	k=8	k=16
Vanilla PM MSR (non-systematic)	50	75	87
Sparse PM MSR (non-systematic)	64	85	93
Vanilla PM MSR (systematic)	0	0	0
Sparse PM MSR (systematic)	50	75	88

Interestingly, such construction based on dense encoding matrix Ψ leads to sparse generator matrix G as shown on Table 1 (*e.g.*, G contains 75% of zeros for $k = 8$). Sparsity is important as multiplications by zero can be skipped resulting in lower computational costs. However, when the code is turned into systematic form, sparsity is lost and the lower part of the resulting generator G'' contains 0 % of zero.

In order to reduce the computational cost of product-matrix MSR codes, we propose an alternative construction which satisfies the conditions aforementioned but gives slightly sparser Ψ and G , and much sparser systematic generator G'' . We take Φ as an identity matrix concatenated to a Cauchy-Matrix (*i.e.*, $\Phi_{i,j} = (g^{i+\alpha} - g^{j-1})^{-1}$ for all $i > \alpha$). Λ is set to $\Lambda_{i,i} = \frac{g^{i+\alpha} - g^0}{g^{i+\alpha} - g^\alpha}$ and $\Lambda_{i,j} = 0$ if $i \neq j$. For $n = 5, k = 3, d = 4, \alpha = 2$ and a finite field having a generator element g , this gives the following matrices:

$$\Phi = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{g^{2\alpha+1}-g^0} & \frac{1}{g^{2\alpha+1}-g^1} \\ \frac{1}{g^{2\alpha+2}-g^0} & \frac{1}{g^{2\alpha+2}-g^1} \\ \frac{1}{g^{2\alpha+3}-g^0} & \frac{1}{g^{2\alpha+3}-g^1} \end{pmatrix} \quad \Lambda = \begin{pmatrix} \frac{g^{\alpha+1}-g^0}{g^{\alpha+1}-g^\alpha} & 0 \\ \frac{g^{\alpha+2}-g^0}{g^{\alpha+2}-g^\alpha} & 0 \\ \frac{g^{2\alpha+1}-g^0}{g^{2\alpha+1}-g^\alpha} & 0 \\ \frac{g^{2\alpha+2}-g^0}{g^{2\alpha+2}-g^\alpha} & 0 \\ \frac{g^{2\alpha+3}-g^0}{g^{2\alpha+3}-g^\alpha} & 0 \end{pmatrix}$$

$$\Psi = [\Phi \quad \Lambda\Phi]$$

By construction, any k rows of Φ are invertible; and for a given finite field and k , it is easy to check computationally that all n values of Λ are different and that all n possible $d \times d$ sub-matrices of Ψ are invertible. The construction we give is valid for $k \in \{2 \dots 39\}$ in the default $GF(2^8)$ from GF-Complete [19] and $k \in \{2 \dots 64\}$ in $GF(2^{16})$. Notice that the number of matrices to check computationally is small so that checking all the conditions for all k takes 0.6 seconds in $GF(2^8)$ and 19 seconds for checking up to $k = 64$ in $GF(2^{16})$.

As shown in Table 1, the non-systematic codes are slightly sparser than the non-systematic vanilla codes (*e.g.*, 85 % instead of 75 % of zeros for $k = 8$). However, the main gain comes for systematic codes, which are significantly sparser than systematic vanilla codes (*e.g.*, 75 % instead of 0 % for $k = 8$) thus leading to faster computation. This is important since most codes deployed are systematic.

An important advantage of the sparser codes concerns repair. The matrices used for repair (lines of Φ) contains

some zeros (instead of none for vanilla codes). Hence, it implies that helper nodes only have to read the part of their data to be multiplied by non-zero coefficients (*i.e.*, 1 block for a failure of the α first nodes, and α blocks otherwise) instead of all their data (*i.e.*, α blocks). Hence, the disk bandwidth of non-failed devices is preserved. The reduction is on average $\frac{k-2}{2k-2}$. For $k = 8$ (resp. $k = 4$), the disk bandwidth is reduced by 43% (resp. 33%). For large values of k , it approaches 50%.

4 Evaluation

In order to evaluate the computational cost of product-matrix regenerating codes, we implemented them using GF-Complete [19] (for the specific version) and Jerasure 2.0 [20] (for the linearized version). We run our benchmark on a Xeon E5-2640 (2.5 Ghz), which supports SSE (up to 4.2) and AVX instruction sets. Hence, as PM codes require small finite fields (*e.g.*, $GF(2^8)$ is sufficient for $k \leq 32$), the libraries can leverage the split-table technique [1, 15] relying on SIMD for efficient multiplications. All benchmarks correspond to single threaded computations (*i.e.*, use a single core) and are averaged over 1,000 runs.

For each operation, we separated the total running time into (i) initialization time (*e.g.*, building generator or encoding matrix, inverting it) and (ii) time needed to apply the operation to the data (reported as the corresponding throughput³ in MB/s). This distinction is important, as in a practical deployment, depending on the operation, the initialization phase could be precomputed and its result stored once for all (*e.g.*, encoding), or must be computed at for each operation as it depends on the failure pattern (*e.g.*, decoding) but can be reused across stripes for large objects.

We evaluate the following codes:

- Vanilla Specific Product-Matrix (PM Spec. Van.) are the codes described in the seminal paper [23].
- Sparse Specific Product-Matrix (PM Spec. Spa.) are the sparse codes described in Section 3.3 using specific algorithms of the seminal paper [23].
- Vanilla Linear Product-Matrix (PM Lin. Van.) are the linearized (see Section 3.1) version of codes described in the seminal paper [23].
- Sparse Linear Product-Matrix (PM Lin. Spa.) are the linearized version (see Section 3.1) of sparse codes described in Section 3.3.
- Reed-Solomon codes (RS Vandermonde) are Vandermonde-based Reed-Solomon codes implemented in Jerasure and are used as a baseline for the performance evaluation.

³The encoding/decoding throughput is the time divided by the amount of data to encode/decode (*i.e.*, $k\alpha$). The repair throughput is the time divided by the amount of data to repair (*i.e.*, α).

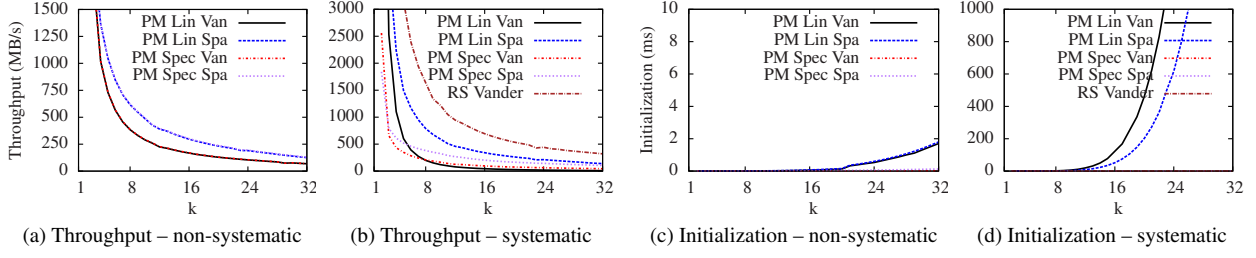


Figure 2: Encoding (regular or systematic) MSR Product-Matrix codes and Reed-Solomon codes.

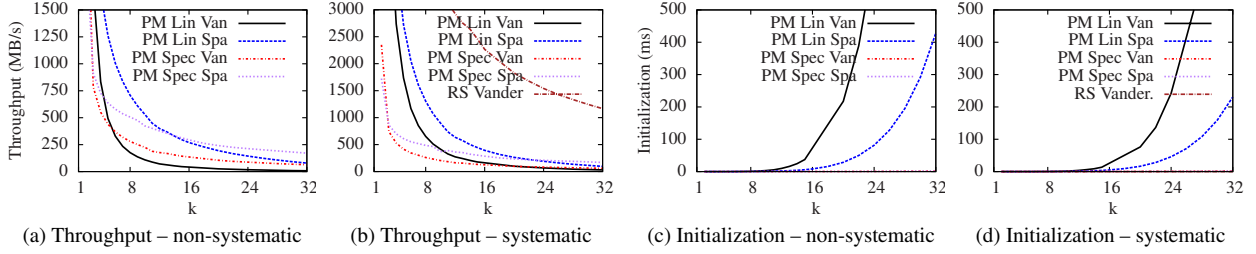


Figure 3: Decoding (regular or systematic) MSR Product-Matrix codes and Reed-Solomon codes.

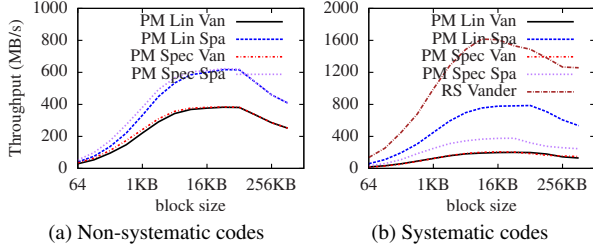


Figure 4: Impact of block size on MSR encoding.

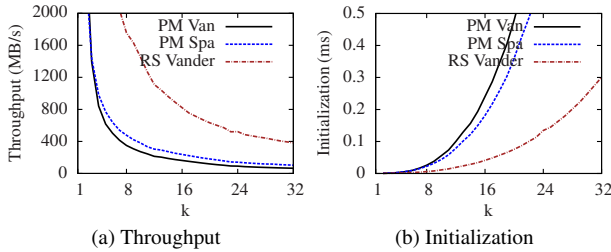


Figure 5: Repair of MSR PM codes and RS codes.

4.1 MSR codes

Figure 4 shows that a block size of 16 KB gives the best performance for encoding. This is coherent with the size of the L1 cache of the processor, which is 32 KB. Similar results can be observed for decoding, repair and MBR codes. Hence, all the subsequent experiments use 16 KB as the block size.

Figure 2 shows the performance of encoding operations. For $k = 8$, a systematic, sparse and linearized PM MSR code allows encoding at 790 MB/s, while Reed-Solomon codes achieve 1640 MB/s and the systematic vanilla PM codes achieve 210 MB/s. Hence, optimized product-matrix codes significantly improve performance over vanilla product-matrix codes; They are a reasonable alternative to Reed-Solomon for practical deployment.

To achieve these speeds, linearizing the code is useful for systematic codes, as it allows encoding directly without performing the *pre*-decoding step thus doubling the throughput (see PM Lin. Spa. vs PM Spec Spa. on Fig. 2b). However, the linearized version is less efficient than the specific version for large values of k when using the vanilla code (see PM Lin. Van. vs PM Spec Van. on Fig. 2b). This is due to the fact that the generator matrix is not sparse for the systematic vanilla code. Hence, the sparsified version presented in this paper is an important optimization allowing to better leverage the gains from linearization. Figure 2d shows that the gain observed in throughput for using linearized systematic product-matrix codes comes at the price of a higher initialization cost (*i.e.*, time needed to compute the systematic generator matrix). Yet, since the generator matrix is constant, it can be computed once and reused for all other encodings.

Figure 5 shows the performance of the repair operation for a single failure. In order to measure the performance, the computation which consist of several independent sub-computations, is performed sequentially on a single core. Sparse PM codes are slightly faster than

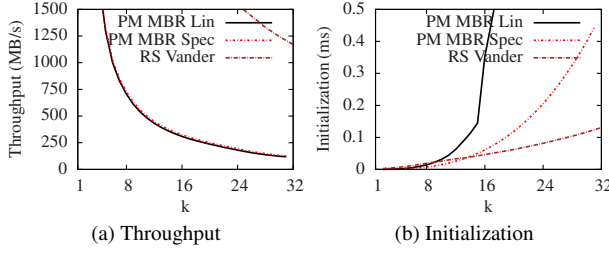


Figure 6: Encoding of MBR PM codes and RS codes.

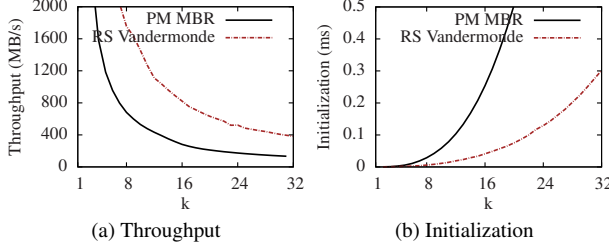


Figure 7: Repair of MBR PM codes and RS codes.

vanilla codes and are operating at 480 MB/s when Reed-Solomon codes operate at 1830 MB/s.

Figure 3 shows the performance of decoding (averaged over many failure patterns with one to $n - k$ failures) using data from systematic devices if available. The initialization cost does not include the generation of the generator matrix, which correspond to the initialization of the encoding operation and can be precomputed once for all. The sparse PM are faster than the vanilla PM. Using the linearized version of sparse PM codes is more interesting for low values of k ($k < 12$ for non-systematic codes, $k < 21$ for systematic codes), and for large data (due to the initialization cost). Notice that decoding method is independent of the encoding method: it is possible to encode using the linear algorithm and to decode using the specific algorithm of the seminal paper [23].

4.2 MBR codes

In this sub-section, we evaluate the computational performance of PM MBR codes [23] which are the most versatile minimum bandwidth regenerating codes known. MBR codes have a higher storage overhead than the corresponding MSR codes but lower the network bandwidth during repair. Hence, they are of interest in context where network is a more scarce resource than storage.

Figure 6 shows the encoding performance of the PM MBR codes, which are naturally sparse and systematic. The PM MBR codes achieve 725 MB/s for $k = 8$, thus having a limited overhead when compared to Reed-Solomon codes (1640 MB/s). This speed is comparable to systematic sparse PM MSR codes.

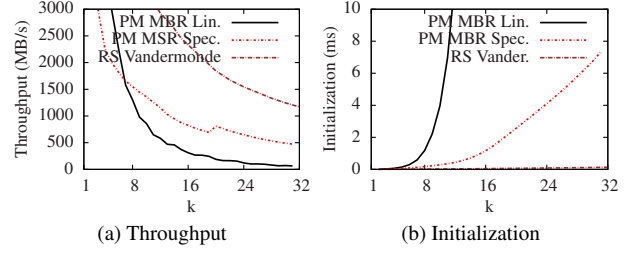


Figure 8: Decoding of MBR PM codes and RS codes.

Figure 7 shows the repair performance of the PM MBR codes. The PM MBR codes achieve 681 MB/s for $k=8$ which is slightly faster than the PM MSR codes and remain reasonable when compared to the 1830 MB/s of Reed-Solomon codes.

Figure 8 shows the decoding performance of the PM MBR codes. In this case, the linearized version is faster only for $k < 7$. Indeed, the specific algorithm for decoding factorizes a lot of computations thus being more efficient. The slight advantage of the linearized version for low values of k comes from the fact that it is possible to selectively decode only a part of the data if only few systematic devices have failed. When k becomes larger, the failure patterns considered (between 1 and $n - k$ failures) have a stronger impact on the systematic devices making this optimization less efficient. Decoding PM MBR codes is faster than all PM MSR: Indeed, the code structure is simpler and the redundancy helps the decoding.

5 Conclusion

We presented new sparse product-matrix regenerating codes that encode systematically at 790 MB/s for typical settings (*i.e.* $k = 8$), four times faster than vanilla product-matrix regenerating codes [23] (210 MB/s). Thanks to this significant improvement, regenerating codes achieve half the throughput of Reed-Solomon codes (1640 MB/s). The achieved performance is the result of a good interaction between the linearization of the systematic code that allows to collapse the pre-processing and the encoding; and the sparse structure of the new product-matrix regenerating codes we define. Additionally, sparse codes lower the impact of repair on disks for non-failed devices (43% less reads for $k = 8$).

Beside this new code, we also report numerous throughputs for existing systematic and non-systematic MSR codes and MBR codes, for decoding and repairing. Throughputs reported give insight on the CPU penalty of using regenerating codes and show that it remains limited when compared to Reed-Solomon codes. Throughput reported are an order of magnitude higher than previous studies, thus highlighting that regenerating codes can be used in practical system with little CPU penalty.

References

- [1] ANVIN, H. P. The mathematics of RAID-6. <https://www.kernel.org/pub/linux/kernel/people/hpa/raid6.pdf>, 2011.
- [2] CADAMBE, V. R., JAFAR, S. A., MALEKI, H., RAMCHANDRAN, K., AND SUH, C. Asymptotic Interference Alignment for Optimal Repair of MDS codes in Distributed Storage. *IEEE Transactions On Information Theory* (2013).
- [3] CALDER, B., WANG, J., OGUS, A., NILAKANTAN, N., SKJOLSVOLD, A., MCKELVIE, S., XU, Y., SRIVASTAV, S., WU, J., SIMITCI, H., HARI-DAS, J., UDDARAJU, C., KHATRI, H., EDWARDS, A., BEDEKAR, V., MAINALI, S., ABBASI, R., AGARWAL, A., UL HAQ, M. F., UL HAQ, M. I., BHARDWAJ, D., DAYANAND, S., ADUSUMILLI, A., MCNETT, M., SANKARAN, S., MANIVANNAN, K., AND RIGAS, L. Windows Azure Storage: a highly available cloud storage service with strong consistency. In *SOSP* (2011).
- [4] CHEN, H. C. H., HU, Y., LEE, P. P. C., AND TANG, Y. NCcloud: Applying Network Coding for the Storage Repair in a Cloud-of-Clouds. *IEEE Transactions on Computers* 63 (2014), 31–44.
- [5] DIMAKIS, A. G., GODFREY, P. B., WU, Y., WAINWRIGHT, M. O., AND RAMCHANDRAN, K. Network Coding for Distributed Storage Systems. *IEEE Transactions On Information Theory* 56 (2010), 4539–4551.
- [6] DIMAKIS, A. G., RAMCHANDRAN, K., WU, Y., AND SUH, C. A Survey on Network Codes for Distributed Storage. *The Proceedings of the IEEE* 99 (2010), 476–489.
- [7] DUMINUCO, A., AND BIERACK, E. A Practical Study of Regenerating Codes for Peer-to-Peer Backup Systems. In *ICDCS* (2009).
- [8] FIKES, A. Storage Architecture and Challenges. http://research.google.com/university-relations/facultysummit2010/storage_architecture_and_challenges.pdf, July 2010.
- [9] FORD, D., LABELLE, F., POPOVICI, F. I., STOKELY, M., TRUONG, V.-A., BARROSO, L., GRIMES, C., AND QUINLAN, S. Availability in Globally Distributed Storage Systems. In *OSDI* (2010).
- [10] GOPALAN, P., HUANG, C., SIMITCI, H., AND YEKHANIN, S. On the locality of codeword symbols. *IEEE Transactions On Information Theory* 58 (2012), 6925–6934. arXiv:1106.3625.
- [11] HU, Y., YU, C.-M., LI, Y. K., LEE, P. P. C., AND LUI, J. C. S. NCFS: On the Practicality and Extensibility of a Network-Coding-Based Distributed File System. In *NetCod* (2011).
- [12] HUANG, C., CHEN, M., AND LI, J. Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems. In *NCA* (2007).
- [13] JIEKAK, S., KERMARREC, A.-M., LE SCOUARNEC, N., STRAUB, G., AND VAN KEMPEN, A. Regenerating Codes: A System Perspective. *ACM SIGOPS Operating Systems Review* 47 (2013), 23–32.
- [14] KHAN, O., BURNS, R., PLANK, J., PIERCE, W., AND HUANG, C. Rethinking Erasure Codes for Cloud File Systems: Minimizing I/O for Recovery and Degraded Reads. In *FAST* (2012).
- [15] LI, R., LIN, J., AND LEE, P. P. C. CORE: Augmenting Regenerating-Coding-Based Recovery for Single and Concurrent Failures in Distributed Storage Systems. In *MSST* (2013).
- [16] LUO, J., SHRESTHA, M., XU, L., AND PLANK, J. S. Efficient Encoding Schedules for XOR-based Erasure Codes. *IEEE Transactions on Computers* (2013).
- [17] PAPAILIOPOULOS, D. S., AND DIMAKIS, A. G. Locally Repairable Codes. In *ISIT* (2012).
- [18] PLANK, J. S., BUCHSBAUM, A. L., AND VANDER ZANDEN, B. T. Minimum density RAID-6 codes. *ACM Transactions on Storage* 6, 4 (May 2011).
- [19] PLANK, J. S., GREENAN, K., AND MILLER, E. L. Screaming Fast Galois Field Arithmetic Using Intel SIMD Extensions. In *FAST* (2013).
- [20] PLANK, J. S., AND GREENAN, K. M. Jerasure: A Library in C Facilitating Erasure Coding for Storage Applications – Version 2.0. Tech. Rep. UT-EECS-14-721, University of Tennessee, January 2014.
- [21] PLANK, J. S., SCHUMAN, C. D., AND ROBISON, B. D. Heuristics for Optimizing Matrix-Based Erasure Codes for Fault-Tolerant Storage Systems. In *DSN* (2012).

- [22] RASHMI, K. V., SHAH, N. B., GU, D., KUANG, H., BORTHAKUR, D., AND RAMCHANDRAN, K. A Solution to the Network Challenges of Data Recovery in Erasure-coded Distributed Storage Systems: A Study on the Facebook Warehouse Cluster. In *HotStorage* (2013).
- [23] RASHMI, K. V., SHAH, N. B., AND KUMAR, P. V. Optimal Exact-Regenerating Codes for Distributed Storage at the MSR and MBR Points via a Product-Matrix Construction. *IEEE Transactions On Information Theory* 57 (2011), 5227–5239.
- [24] SHAH, N. B., RASHMI, K., KUMAR, P. V., AND RAMCHANDRAN, K. Distributed Storage Codes with Repair-by-Transfer and Non-achievability of Interior Points on the Storage-Bandwidth Trade-off. *IEEE Transactions On Information Theory* 58 (2012), 1837–1852.
- [25] SHAH, N. B., RASHMI, K., KUMAR, P. V., AND RAMCHANDRAN, K. Interference Alignment in Regenerating Codes for Distributed Storage: Necessity and Code Constructions. *IEEE Transactions On Information Theory* 58 (2012), 2134–2158.
- [26] SUH, C., AND RAMCHANDRAN, K. On the Existence of Optimal Exact-Repair MDS Codes for Distributed Storage. *ArXiv e-prints* (2010). arXiv:1004.4663.
- [27] SUH, C., AND RAMCHANDRAN, K. Exact-Repair MDS code construction using interference alignment. *IEEE Transactions On Information Theory* 57 (2011), 1425–1442.
- [28] TAMO, I., WANG, Z., AND BRUCK, J. Access vs. Bandwidth in Codes for Storage. In *ISIT* (2012).
- [29] TAMO, I., WANG, Z., AND BRUCK, J. Zigzag Codes: MDS Array Codes with Optimal Rebuilding. *IEEE Transactions On Information Theory* 59 (2013), 1597–1616.